Am. J. Hum. Genet. 61:1179–1188, 1997

Allele-Sharing Models: LOD Scores and Accurate Linkage Tests

Augustine Kong[1,3] and Nancy J. Cox[2]

Departments of [1]Statistics and [2]Medicine, The University of Chicago, Chicago; and [3]deCODE Genetics, Reykjavik

Summary

Starting with a test statistic for linkage analysis based on allele sharing, we propose an associated one-parameter model. Under general missing-data patterns, this model allows exact calculation of likelihood ratios and LOD scores and has been implemented by a simple modification of existing software. Most important, accurate linkage tests can be performed. Using an example, we show that some previously suggested approaches to handling less than perfectly informative data can be unacceptably conservative. Situations in which this model may not perform well are discussed, and an alternative model that requires additional computations is suggested.

Introduction

Linkage tests based on allele sharing are popular for mapping susceptibility genes for complex traits (Weeks and Lange 1988; Risch 1990*a,* 1990*b;* Whittemore and Halpern 1994). Compared with traditional parametric methods, these approaches have the advantage of not having to specify an inheritance model explicitly. Building on this earlier work, Kruglyak et al. (1996) have proposed a procedure, based on the idea of scoring functions defined with respect to identity by descent (IBD) sharing, that can be applied to an arbitrary mixture of family structures. Most important, the procedure is incorporated into the computer program GENEHUNTER, which performs multipoint calculations and hence allows the full utilization of the information available on descent, the inheritance pattern at every locus. However, the linkage test proposed by Kruglyak et al. is conservative when the descent information is incomplete, which is nearly always the case, because it overestimates the variance of the statistic that it uses. The amount of overestimation is severe when the information on descent is very far from complete, in which case the procedure can be unacceptably conservative. Also, with traditional parametric procedures, when multipoint calculations are performed, the LOD-score function can be used to make comparisons among loci and to construct confidence/support regions for the gene location. The procedure of Kruglyak et al. does not provide that. Following Whittemore (1996), we argue that any choice of the scoring functions and the associated weighting factors has an implicit underlying one-parameter alternative model for both the distribution of the inheritance vector and sharing probabilities. Likelihoods and LOD scores are defined with respect to this allele-sharing model. In particular, we show that the LOD scores can be easily computed for any general missing-data pattern, by use of the standard output of GENEHUNTER. These LOD scores can be used just like the traditional LOD scores, for the purposes of constructing support regions and comparing loci. Also, an accurate likelihood-ratio test is then available for the purpose of assessing the evidence for linkage. As an example, our approach as implemented in a modified version of GENEHUNTER is applied to data derived from a genome scan for non–insulin-dependent diabetes mellitus (NIDDM) susceptibility loci in Mexican Americans (Hanis et al. 1996). Results of multipoint analyses of the chromosome 2 region providing evidence for the NIDDM1 locus reveal that the likelihood-ratio test gives a *P* value more than an order of magnitude smaller than that given by the GENEHUNTER nonparametric-linkage (NPL) analysis. An analysis of a chromosome 2 framework map in which markers are ~20 cM apart illustrates additional advantages of our method. With the framework map, our LOD scores tend to curve upward between markers, as do traditional parametric LOD scores. By contrast, the NPL scores tend to curve downward between markers, a consequence of the lack of adjustment for incomplete information. We also point out that special care is required when data sets consisting of a small number of pedigrees are analyzed, because of the potential breakdown of asymptotic approximations.

Data and Scoring Functions

The data will, in general, consist of *m* pedigrees and genotype data on some markers for some pedigree mem-

Received October 3, 1996; accepted for publication August 19, 1997; electronically published October 29, 1997.

Address for correspondence and reprints: Dr. Augustine Kong, Department of Statistics, The University of Chicago, Chicago, IL 60637. E-mail: kong@galton.uchicago.edu

0002-9297/97/6105-0025$02.00

bers. With complex diseases, each pedigree will usually have two or more affected individuals. The family structure and pattern of affecteds can be quite complex and completely different for different pedigrees. However, because of the computational demands of multipoint calculations, there is a limitation to the size of a pedigree. With GENEHUNTER, the upper bound is ~12 nonfounders (Kruglyak et al. 1996). On the other extreme, the procedure can also be applied to sib-pair data for which each pedigree is a family of two parents and two affected children. Under the null hypothesis (H_0) that a locus is not linked to a disease-susceptibility gene, the statistical behavior of the number of alleles IBD among individuals depends only on their relationships to each other, as determined by the pedigree structure, and not on their disease status. For a locus that is linked to a disease-susceptibility gene, there is expected to be an increase in the number of alleles IBD among the affecteds, relative to null expectation. Testing for linkage becomes testing for excess sharing. The magnitude of the excess will, in general, depend on the mode of inheritance and on the distance between the linked locus and the disease-gene locus. However, with complex diseases, the mode of inheritance is unknown, and it may sometimes be easier to model the degree and the direction of excess directly.

Consider any locus that is not necessarily a marker locus but that has one or more markers in the neighborhood. For pedigree i let S_i be some function that is defined on the basis of IBD sharing at this locus among the affecteds. Following the suggestions of Whittemore and Halpern (1994), Kruglyak et al. (1996) implemented the two scoring functions, $\mathbf{S}_{\text{pairs}}$ and $\mathbf{S}_{\text{all}}$, in GENEHUNTER. $\mathbf{S}_{\text{pairs}}$ is simply the number of pairs of alleles from distinct affected pedigree members that are IBD. In comparison, $\mathbf{S}_{\text{all}}$ puts extra weight on three or more affecteds sharing the same allele IBD. For the exact definition, the reader is referred to the above-mentioned articles. In general, S_i can be any function that has a higher expected value under linkage than under no linkage. In theory, the definition of S_i can also involve allele sharing—or lack of sharing—between affecteds and unaffecteds. Some comments on how to choose S_i will be given below. Here, suppose that the choice of S_i has been made. The standardized form of S_i is defined as

$$Z_i = \frac{S_i - \mu_i}{\sqrt{\sigma_i^2}} = \frac{S_i - \mu_i}{\sigma_i} ,$$

where $\mu_i = E(S_i \mid H_0)$. Note that Z_i has mean 0 and variance 1 under H_0. Consider a linear combination

$$Z = \frac{\sum_{i=1}^{m} \gamma_i Z_i}{\sqrt{\sum_{i=1}^{m} \gamma_i^2}} ,$$

where $\gamma_i \geqslant 0$ are weighting factors. The denominator of Z ensures that Z has variance 1 under H_0. It is obvious that the γ_i are relative in the sense that they are only important up to a multiplicative constant. Indeed, Kruglyak et al. added the constraint $\Sigma\gamma_i^2 = 1$, so that Z reduces to $\Sigma_i\gamma_i Z_i$. We prefer not to have that constraint, to simplify exposition in some discussions. For large m, the distribution of Z under H_0 can be approximated well by the standard normal distribution, and approximate (one-sided) P values can be computed on the basis of that, if Z is directly observed. Since normal approximation can break down with a small sample size, GENEHUNTER, with some additional computations, provides an exact P value by enumerating the distribution of Z under H_0. The optimal choice of the γ_i, from the aspect of maximization of power, will, in general, depend on many factors, including the mode of inheritance. Some comments on the choice of γ_i will be given below.

In general, the information on descent is incomplete, and the Z_i and, hence, Z are not fully determined by the data. However, $\overline{S}_i = E(S_i \mid \text{data}, H_0)$and $\overline{Z}_i = E(Z_i \mid \text{data}, H_0) = (\overline{S}_i - \mu_i)/\sigma_i$ can always be computed from the observed data. (Here we follow the notation of Kruglyak et al., but the reader should be warned that $\overline{S}_i$ and $\overline{Z}_i$ are expectations computed with respect to a distribution and are not sample averages.) With GENEHUNTER, the conditional expectation is computed by use of all marker data available and is, in general, a multipoint calculation. Also note that the expectation is conditional on the null hypothesis of no linkage. If the data do not determine Z_i, then the expectation conditional on an alternative of linkage is different and, in general, higher. (From here on, when it is not explicitly noted otherwise, expectations and variances are under H_0.) Kruglyak et al. propose the statistic

$$\overline{Z} = \frac{\sum_{i=1}^{m} \gamma_i \overline{Z}_i}{\sqrt{\sum_{i=1}^{m} \gamma_i^2}} ,$$

which is referred to as the "NPL score." They note that, under H_0, $E(\overline{Z})$ is still 0 but that the variance of $\overline{Z}_i$ and, hence, the variance of $\overline{Z}$ will, in general, be <1. However, they recommend the continued use of the standard normal distribution—or of the exact distribution of Z—as

references to produce a P value. They call this the "perfect-data approximation" and note that this will, in general, be a conservative procedure.

How conservative the procedure is depends on how imperfect the information on descent is. That information is often far from complete, which can be caused by various combinations of (1) many untyped members, (2) low heterozygosity of markers, and (3) wide spacing of markers. Performing multipoint calculations certainly is an advantage, since it allows the full utilization of the information available. However, since linkage analyses span such a wide spectrum, it is not difficult to find any real examples in which the perfect-data approximation is unacceptably conservative. To eliminate the conservativeness, we propose a likelihood approach.

A One-Parameter Allele-Sharing Model

At a particular location, not necessarily a marker locus, for pedigree i, let v_i denote the inheritance vector that can assume N_i configurations. (v_i contains all information on descent. In particular, the number of alleles distinct by descent is equal to two times the number of founders. For each nonfounder, the inheritance vector determines which two of the founder alleles were inherited.) Under H_0, each configuration has the same probability, $c_i = N_i^{-1}$, and so $\mu_i = \Sigma c_i S_i(w)$ and $\sigma_i^2 = [\sum_w c_i S_i{}^2(w)] - \mu_i{}^2$. (Our notation here uses w to denote a configuration of v_i. Sums are over all the possible configurations of v_i.) Now suppose that a single-parameter alternative model is introduced with δ as the free parameter. Let δ be chosen so that $\delta = 0$ corresponds to H_0 and so that $\delta \geqslant 0$ corresponds to the alternative of excess sharing. For each w, let $P_i(w \mid \delta) = P(v_i = w \mid \delta)$ be the probability specified by the model. If the v_i can be directly observed, then the log likelihood is simply $l(\delta) = \sum_{i=1}^{m} \ln [P_i(v_i \mid \delta)]$. We can then find the maximum-likelihood estimate of δ and use the likelihood-ratio χ^2 statistic to test $H_0{:}\delta = 0$. However, our information on descent is, in general, incomplete, and v_i cannot be determined by the data. Indeed, suppose that we are restricted to procedures that use only the values $\overline{Z}_i$ computed by GENEHUNTER. Remarkably, there exists a class of models for which the log likelihood $l(\delta) = \ln[P(\text{all data} \mid \delta)]$ can be written on the basis of only the $\overline{Z}_i$. The probabilities of v_i are in the form

$$\begin{aligned} P_i(w|\delta) &= P(v_i = w|\delta) \qquad (1)\\ &= P(v_i = w|H_0)\left\{1 + \frac{\delta\gamma_i[S_i(w) - \mu_i]}{\sigma_i}\right\}\\ &= c_i\left\{1 + \frac{\delta\gamma_i[S_i(w) - \mu_i]}{\sigma_i}\right\}, \end{aligned}$$

and the γ_i are weighting factors that in theory can be chosen on the basis of the pedigree structures. With this model, it is proved in appendix A that the log likelihood, for the sums of the m pedigrees, is

$$\begin{aligned} l(\delta) &= C + \sum_{i=1}^{m} \ln[1 + \delta\gamma_i(\overline{S}_i - \mu_i)/\sigma_i]\\ &= C + \sum_{i=1}^{m} \ln\left(1 + \delta\gamma_i\overline{Z}_i\right)\\ &= C + \sum_{i=1}^{m} \ln(1 + \delta T_i), \end{aligned}$$

where $T_i = \gamma_i\overline{Z}_i$ and C is the constant that depends on the data but not on δ. In contrast with equation (1), here $S_i(v_i)$ is replaced by $\overline{S}_i$, and one might think that we are simply providing an approximation to the log likelihood, but that is not the case. We emphasize that, with any missing-data pattern, this is the exact log likelihood, which of course has to coincide with the complete-data log likelihood when v_i happens to be determined by the data. Other properties of model (1), including its relationship with the test based on the NPL score, will be discussed later. Here, suppose that we accept it as an appropriate model. Let a_i be the smallest theoretically possible value of S_i. For $P(S_i = a_i)$ not to be assigned a negative value, δ is bounded above by $b_i = \sigma_i/[\gamma_i(\mu_i - a_i)]$. This implies that the legitimate range of δ is between 0 and $b = \min_i b_i$.

Let $0 \leqslant \hat{\delta} \leqslant b$ be the maximum-likelihood estimate of δ. Then

$$2[l(\hat{\delta}) - l(0)] = 2\sum_{i=1}^{m} \ln(1 + \hat{\delta}T_i)$$

is the χ^2 statistic with 1 df for testing $H_0{:}\delta = 0$. Since our test is one sided ($\delta > 0$), define

$$Z_{1r} = \sqrt{2\left[l\left(\hat{\delta}\right) - l\left(0\right)\right]}.$$

With large N, when $Z_{1r} > 0 (\hat{\delta} > 0)$, the P value can be approximated by $1 - \Phi(Z_{1r})$, where Φ is the cumulative distribution of the standard normal distribution.

So far, we have suppressed the role of the location. For a location x, let $l(x,\delta)$ be the joint log likelihood—that is, the log of the probability of the data computed under the assumption that x is the gene location and δ is the amount of deviation of the distribution of v_i from its null distribution, a measure of the gene effect. Let $\hat{\delta}_x$ be the maximum-likelihood estimate of δ conditional on x; then

$$l(x, \hat{\delta}_x) - l(x, 0) = \sum_{i=1}^{m} \ln[1 + \hat{\delta}_x T_i(x)] \ .$$

For any two loci x and y, $l(x, 0) = l(y, 0)$, since, if a gene has no effect, it does not matter where it is. (Note that this is only true if the same data are used to calculate both, and hence multipoint calculation is of utmost importance.) So

$$\begin{aligned} l(x, \hat{\delta}_x) - l(y, \hat{\delta}_y) &= [l(x, \hat{\delta}_x) - l(x, 0)] \\ &\quad -[l(y, \hat{\delta}_y) - l(y, 0)] \\ &= \sum_{i=1}^{m} \ln[1 + \hat{\delta}_x T_i(x)] \\ &\quad - \sum_{i=1}^{m} \ln[1 + \hat{\delta}_y T_i(y)] \ . \end{aligned}$$

Hence

$$\begin{aligned} \mathrm{LOD}^*(x) &= \frac{\sum_{i=1}^{m} \ln[1 + \hat{\delta}_x T_i(x)]}{\ln(10)} \\ &\approx \frac{\sum_{i=1}^{m} \ln[1 + \hat{\delta}_x T_i(x)]}{2.3} \end{aligned}$$

is the LOD score with respect to model (1). Clearly, $Z_{\mathrm{lr}}(x) = \sqrt{\mathrm{LOD}^*(x) \times 2\ln(10)}$ and $\mathrm{LOD}^*(x) = Z_{\mathrm{lr}}^2(x)/2\ln(10)$, where $Z_{\mathrm{lr}}(x)$ denotes Z_{lr} evaluated at location x.

Our LOD score $\mathrm{LOD}^*(x)$ is very similar to the LOD scores obtained in a traditional parametric analysis (e.g., in fitting a dominant model). One slight difference is that in traditional parametric analysis the LOD score as a function of x is usually plotted with the parameters (i.e., penetrances and disease-allele frequency) having the same values for all the locations. In our case, $\hat{\delta}_x$ is, in general, different at different locations and $\mathrm{LOD}^*(x)$ is always non-negative. A situation in which one may consider calculating $\mathrm{LOD}^*(x)$ with a fixed value of δ, in which case $\mathrm{LOD}^*(x)$ can be negative, is *exclusion mapping*. However, more work has to be done to ensure that results from such analyses can be appropriately interpreted.

Example

In order to illustrate the differences between the results that would be obtained by use of the NPL score of Kruglyak et al. (1996) and those that would be obtained by the test statistic Z_{lr} that we propose, we have modified GENEHUNTER to calculate LOD* and Z_{lr}, in addition to the NPL score $\overline{Z}$. Both a brief description of this modified version of GENEHUNTER and information on obtaining it via an anonymous ftp site are provided in appendix B. As an example data set, we have analyzed the chromosome 2 data first reported, by Hanis et al. (1996), in a genomewide scan for NIDDM susceptibility loci in Mexican Americans. The data include 170 sibships each having at least two affected sibs but no parents and no unaffected sibs and are described in more detail in the original article. The original full chromosome 2 map included 50 markers, but many regions were densely mapped because (1) there was insufficient information with the initial map, (2) there was some evidence for linkage with the initial map, or (3) there was a candidate gene(s) in the region. Therefore, in addition to the full chromosome 2 map, to mimic what is usually encountered in the first stage of a genomewide scan, we have constructed a framework map, using a subset of the original markers. There are 16 loci in the framework map, which has a resolution of ~20 cM. Figure 1 summarizes the results of the analyses of the full map (fig. 1*a–c*) and the framework map (fig. 1*d–f*) for the chromosome 2 data. In these analyses, we report results from use of the scoring function $\mathbf{S}_{\mathrm{all}}$ with equal weighting (γ_i = constant), noting that results obtained from use of $\mathbf{S}_{\mathrm{pairs}}$ differ only trivially from these. And, although all markers were used in the multipoint calculations, the figures focus on the 2qter region where the most significant results are observed. In figure 1, for the two maps, respectively, panels *a* and *d* compare the NPL score $\overline{Z}$ and Z_{lr}; panels *b* and *e* make comparisons in the LOD-score scale, with the dotted line corresponding to $\overline{Z}^2/2\ln(10)$; and panels *c* and *f* display the information statistic of Kruglyak et al. Although there were no parents available for genotyping, the markers are sufficiently polymorphic that the information statistic rarely is <.5, even for the framework map. However, even in the full map, the information statistic rarely is >.8. With the full map, the NPL score and Z_{lr} are quite comparable in regions for which there is little evidence for linkage. But, as the evidence for linkage increases, the conservativeness of the NPL score becomes apparent. Indeed, when the *P* values from the NPL score are used, the 2qter location having the strongest evidence for linkage when the full map is used does not achieve a level of significance sufficient, in the context of a full genome scan, to establish a susceptibility locus ($\overline{Z} = 3.75$, $P = 8.8 \times 10^{-5} > 2 \times 10^{-5}$). However, at the same location, the *P* value associated with the Z_{lr} statistic is comfortably within the guidelines suggested by Lander and Kruglyak (1995) ($\mathrm{LOD}^* = 4.16$, $Z_{\mathrm{lr}} = 4.37$, $P = 6.2 \times 10^{-6} < 2 \times 10^{-5}$). (For simplicity, the *P* values reported here are all based on normal approximation. However, we have used GENEHUNTER to check that, for these families, the exact and normal approximation *P* values are

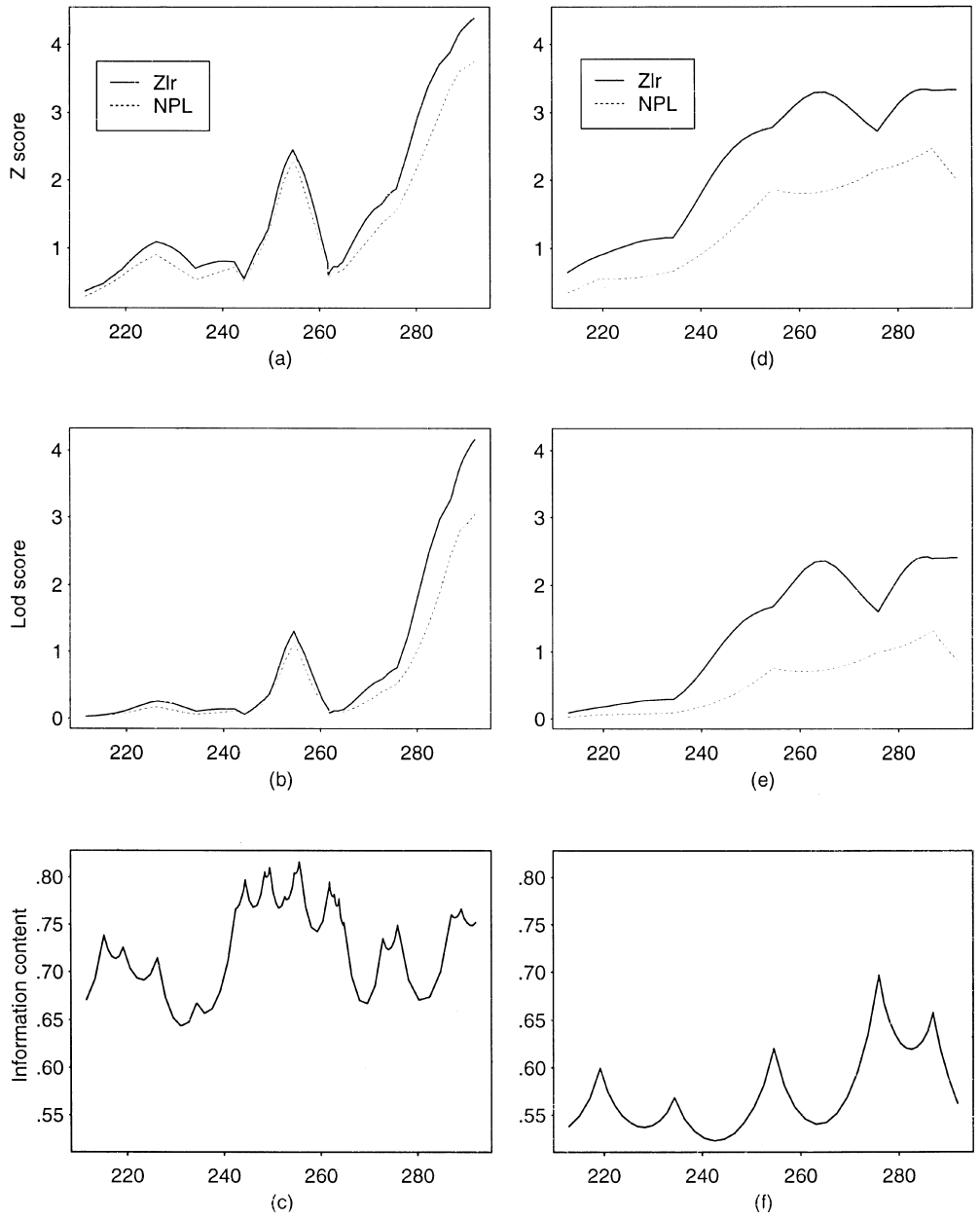

Figure 1 Linkage analysis of NIDDM, for chromosome 2. Shown are results from the full marker map (50 markers) (*a–c,*) and results from the framework map (16 markers, ~20 cM between adjacent markers) (*d–f*). These are multipoint calculations using all the markers simultaneously, although the results presented focus on the 2qter region. The units of the *x*-axis are centimorgans. *a* and *d*, Z_{lr} and NPL score $\overline{Z}$, plotted for comparison. *b* and *e*, Comparisons made under the LOD-score scale. The two curves are LOD* and $\overline{Z}^2/2\ln(10)$. *c* and *f*, Information content. Note that the peaks correspond to marker locations.

very close for *P* values in this range, which is not surprising, since the number of families is large. As will be discussed below, the use of normal approximation can be quite problematic when the data set consists of only a small number of pedigrees.) Note that, although the ratio between Z_{lr} and $\overline{Z}$ is only (4.37/3.75) = 1.17, because of the tail behavior of the normal distribution, the ratio of the *P* values is ~14. With the framework map, because the information is far from complete, $\overline{Z}$ is substantially lower than Z_{lr} everywhere. The biggest difference occurs at location 263 cM, where $\overline{Z} = 1.83$ ($P = .033$) and $Z_{lr} = 3.29$ (LOD* = 2.35, $P = 5.0 \times 10^{-4}$). The ratio of the *P* values is 69. Moreover, as seen in panels *d* and *e* of figure 1, even the general shapes of the curves confirm that the NPL score is not fully utilizing the information available, since the NPL curves

trough between markers. This is due to the fact that the information is more incomplete for a location midway between two markers than it is for a location close to a marker and that no adjustment for incomplete information is taken. In contrast, the Z_{lr} and LOD* curves conform to the more traditional curves obtained for parametric linkage analyses. Most important, with the framework map, the NPL score is sufficiently conservative that this region might not be assigned much priority for the additional follow-up that could greatly increase the evidence for linkage. The additional calculations did not add detectably to the computational time in GENEHUNTER.

More on the Model

A model in the form of model (1) can be found in a report by Whittemore (1996), although there the model is presented in a setting where the pedigrees are assumed to have identical structures and phenotype patterns. Also, the model there allows for more than one parameter, an issue on which we will comment below, in the Additional Parameters section. Following the results in Whittemore, one can easily see that

$$l'(0) = \sum_{i=1}^{m} T_i = \sum_{i=1}^{m} \gamma_i \overline{Z}_i$$

and that the classical *score statistic* (Cox and Hinkley 1994) is

$$\frac{l'(0)}{\sqrt{E[-l''(0)|H_0]}} = \frac{\sum_{i=1}^{m} T_i}{\sqrt{\text{Var}\left[\sum_{i=1}^{m} T_i\right]}} = \frac{\sum_{i=1}^{m} \gamma_i \overline{Z}_i}{\sqrt{\sum_{i=1}^{m} \gamma_i^2 \text{Var}[\overline{Z}_i]}} .$$

With perfect data, $\overline{Z}_i = Z_i$ and the score statistic is $Z = \sum_i \gamma_i Z_i / \sqrt{\sum_i \gamma_i^2}$. For regular statistical problems, the score statistic has a standard normal distribution asymptotically ($m \to \infty$), and the test based on it is asymptotically equivalent to the test based on the likelihood-ratio statistic. For example, in our case, if m is large and $\hat{\delta} > 0$ but not very large, then the observed value of Z_{lr} is expected to be close to the observed value of the score statistic. Sometimes the score statistic is used instead of the likelihood-ratio statistic, because it does not require the evaluation of $\hat{\delta}$ and is hence computationally simpler. Indeed, with the application here, if the data were complete, then Z would be slightly easier to compute than Z_{lr}. However, the situation is entirely different with incomplete data. Here, Z_{lr} can still be easily computed. By contrast, the evaluation of the score statistic requires that multipoint simulation be performed, in order to determine $\text{Var}[\overline{Z}_i]$. Approximating $\text{Var}[\overline{Z}_i]$ by $\text{Var}[Z_i] = 1$ (perfect-data approximation) can, as we have demonstrated, lead to very conservative results.

For pedigree i, let $\mu_{i\delta} = E(S_i \,|\delta)$ be the mean of S_i under the alternative model specified by model (1). So, $\mu_i = \mu_{i0}$, and

$$\begin{aligned}
\mu_{i\delta} &= \mu_i + (\delta\gamma_i/\sigma_i) \sum_w c_i S_i(w)[S_i(w) - \mu_i] \\
&= \mu_i + (\delta\gamma_i/\sigma_i) \left\{ \sum_w c_i S_i(w) S_i(w) - \mu_i \sum_w c_i S_i(w) \right\} \\
&= \mu_i + (\delta\gamma_i/\sigma_i) \{E[S_i^2(w)] - \mu_i^2\} \\
&= \mu_i + (\delta\gamma_i/\sigma_i)\sigma_i^2 \\
&= \mu_i + \delta\gamma_i\sigma_i .
\end{aligned}$$

In other words, for the various S_i, when their means deviate from the null, model (1) specifies that the deviations are proportional to $\sigma_i\gamma_i$. Note that the model does not specify the exact deviation of the mean of an individual S_i, because δ is a free parameter; it does, however, specify the *relative* sizes of the deviations. Notice that $E(Z_i \,|\delta) = \delta\gamma_i$, so the deviations of the standardized forms are proportional to γ_i, the weighting factors. Hence, in vector form, we can think of $\langle\gamma\sigma\rangle$ and $\langle\gamma\rangle$ as the *directions* of deviations of the means of S_i and Z_i, respectively. The test using Z_{lr} will be fully efficient (most powerful) if this is the actual direction of deviation. We note that the test is still valid even if the true deviations do not satisfy the constraint specified by model (1). The test is *model free* in that sense. The penalty will be some loss of power.

For a particular pedigree i, model (1) specifies that

$$\frac{P[S_i(w) = k|\delta]}{P[S_i(w) = k|H_0]} = 1 + (\delta\gamma_i/\sigma_i)(k - \mu_i) ,$$

so the relative change of $P[S_i(w) = k]$ is proportional to $k - \mu_i$. Consider affected (full) sib pairs with $S_i = S_{pairs}$. (We note that, for affected sibships that have either two or three affected sibs with the parents classified as unknowns or unaffecteds, S_{pairs} and S_{all} are equivalent after standardization.) Here $a_i = 0$, $\mu_i = 1$, and $\sigma_i = \sqrt{1/2}$. Under model (1) , the probabilities of sharing 0, 1, and 2 alleles are, respectively, $(1/4)[1 - (\delta\gamma_i/\sigma_i)]$, (1/2), and $(1/4)[1 + (\delta\gamma_i/\sigma_i)]$. The chance of sharing one allele does not change, because the expected number of alleles shared is 1 under H_0. As noted by Whittemore (1996), this corresponds to the additive genetic model. Maximum deviation corresponds to $\delta = b_i = \sigma_i/\gamma_i$ and to sharing probabilities 0, 1/2, and 1/2, respectively, corre-

sponding to IBD sharing of 75%. In general, if the data set consists entirely of affected sib pairs and the γ_i are set to 1, then $\mu_{i\delta}/2 = (1/2) + (\delta\sigma_i/2)$ is the proportion of alleles IBD in the population of affected sibs ($\sigma_i = \sqrt{1/2}$). So $(1/2) + (\hat{\delta}_x\sigma_i/2)$ is the maximum-likelihood estimate of the proportion of alleles shared under the additive model at location x. We understand that this estimate is now implemented in the program SIBPAL of the SAGE (1994) package (R. Elston, personal communication). With missing data, the previous estimate of the proportion of alleles IBD that is given by SIBPAL is negatively biased with incomplete information.

In general, because of the bound on δ, model (1) does restrict the amount of possible deviation. This restriction is irrelevant when the apparent excess sharing is modest, as is expected for most data on complex diseases. (It is to be noted that, even with modest excess sharing, very significant results can be obtained with large sample sizes, as can be seen with the Hanis et al. data.) The bound on δ becomes relevant when extreme excess sharing is observed in a data set with a small number of pedigrees.

Small Number of Families

With a large number of pedigrees, the P value can be well approximated by applying normal approximation to Z_{lr}. Normal approximation may not work well when the data set consists of only a small number of pedigrees and very high sharing is observed. In this case, applying normal approximation to Z_{lr} can give a very conservative P value. The reason is that δ has an upper bound that does not allow model (1) to represent very excessive sharing. When $\hat{\delta}$ is equal or very close to the upper bound of δ, normal approximation is unreliable. As an illustration, consider a single pedigree with five affected sibs. Suppose that all five sibs have inherited the same allele IBD from each of the parents and that the information is complete. The chance of this under H_0, which is also the exact P value, is $(1/2)^8 = .0039$. For scoring functions $\mathbf{S}_{\text{pairs}}$ and $\mathbf{S}_{\text{all}}$, Z_{lr} is 1.84 and 1.92, respectively, and in both cases the likelihood is maximized at the upper bound of δ (with $\gamma = 1, \hat{\delta} = b = 1$ for both scoring functions). Normal approximation gives P values of .033 and .027, respectively, which are much too large. It is interesting that this same example also illustrates the problem of applying normal approximation to the NPL score, but the breakdown is in another direction. For $\mathbf{S}_{\text{pairs}}$ and $\mathbf{S}_{\text{all}}$, the NPL score Z is 4.472 and 5.314, respectively. Normal approximation gives P values of 3.9×10^{-6} and 5.4×10^{-8}, respectively, which are orders of magnitude too small. This is because the distribution of Z, with either scoring function but particularly with $\mathbf{S}_{\text{all}}$, has a very long right tail that is not approximated well by the standard normal distribution. Although this example is an extreme case, a similar phenomenon can be seen with data consisting of a modest number of nuclear pedigrees with very high observed sharing. The P value computed by GENEHUNTER on the basis of the exact distribution of Z is designed for such circumstances. The problem is that it is conservative when there is missing information. With Z_{lr}, because the overall bound on δ is the minimum of the bounds of the individual pedigrees, the problem can be particularly serious when the pedigrees are of very different structures and sizes. One should be aware that the P value obtained by normal approximation is probably conservative when $\hat{\delta}$ is equal or very close to the bound.

When very high sharing is observed in a small number of pedigrees and the information is far from complete, getting a good approximation of the P value without extensive simulation is difficult. Here we give some preliminary results of an approach that we are currently researching. Instead of model (1), consider the model

$$P(v_i = w|\delta) = P(v_i = w|H_0)r_i(\delta)\exp\left\{\frac{\delta\gamma_i[S_i(w) - \mu_i]}{\sigma_i}\right\}$$

$$= c_i r_i(\delta)\exp\left\{\frac{\delta\gamma_i[S_i(w) - \mu_i]}{\sigma_i}\right\},$$

where

$$r_i(\delta) = \left(c_i\sum_w \exp\left\{\frac{\delta\gamma_i[S_i(w) - \mu_i]}{\sigma_i}\right\}\right)^{-1}$$

is the renormalization constant that ensures that $\Sigma_w P(v_i = w \mid \delta) = 1$. $Z_{\text{lr}}(x)$ and LOD*(x) can be similarly defined. We will call this the exponential model and will call model (1) the linear model. When δ is small, $\exp\{\delta\gamma_i[S_i(w) - \mu_i]/\sigma_i\}$ is approximately $1 + \{\delta\gamma_i[S_i(w) - \mu_i]/\sigma_i\}$, and the two models are very close. Indeed, it can be shown that the score statistic corresponding to the exponential model is exactly the same as the score statistic of the linear model that we gave earlier. The exponential model has several nice properties not shared by the linear model. First, with complete data, the NPL score Z is the sufficient statistic and Z_{lr} is a monotonic function of Z. Hence the test based on Z is equivalent to the test based on Z_{lr}, without having to appeal to asymptotics. Most important, with the exponential model, δ does not have an upper bound. When $\delta \to \infty$, the probabilities are concentrated on w with $S_i(w)$ achieving the maximum possible value of S_i. For example, with sib pairs, $P(S_i = 2) \to 1$ as $\delta \to \infty$. However, the exponential model lacks the special missing-data property of the linear model. In particular, with missing data, $l(\delta)$ cannot be written down just on the basis of the conditional expectations $\overline{Z}_i$. Instead it requires the entire conditional distributions of the Z_i.

Hence, evaluating $l(\delta)$ and Z_{lr} for the exponential model is computationally more intensive, but it is not insurmountable. We expect to have a computer program ready for distribution in the very near future. With a working program, we have computed Z_{lr} for the Hanis et al. data by using the exponential model, and the results are very similar to those computed by use of the linear model. This is expected, because of the large number of families.

In contrast to the linear model, the exponential model can provide a very good fit to data consisting of a small number of pedigrees with very extreme IBD sharing. However, the application of normal approximation to Z_{lr} can remain problematic. Consider the previous example of the five affected sibs with perfect IBD sharing. With either scoring function, $Z_{lr} = 3.33$, and normal approximation gives $P = .00043$. This P value is too small, although it is much better than those obtained by application of normal approximation to the NPL scores. This is because the distribution of Z_{lr} is skewed, but much less so than the distribution of Z. We are currently working on methods that are based on the exponential model and that can provide suitable adjustments to the P-value approximations.

Additional Parameters

We have so far considered a single-parameter alternative model. It is, however, quite easy to introduce extra parameters into the model. For example, there may be two types of affected relative pairs (or the same type of relatives but with data collected from two different populations). Specifying the weights corresponds to specifying the ratio of excess sharing. An alternative is to set γ_i to be 1 for one type of affected pairs and to let γ_i of the other type of affected pairs be a free parameter. Together with δ, we will now have two free parameters in the maximization. Of course, if that is done, the associated df of the likelihood-ratio test will increase to 2. (Actually, with a one-sided test, asymptotically, the P value can be approximated by one-half of the tail area of a χ^2 distribution with 1 df plus one-quarter of the tail area of a χ^2 distribution with 2 df.) In general, even with completely different pedigrees, one can think of creating several groups of pedigrees on the basis of similarities in the pedigree structure and in the number of affecteds, and each group can have its own γ treated as a parameter to be fitted. In another direction, when there are covariates, one can have a model relating γ_i to the covariates, with parameters to be fitted. Extra parameters can also be introduced through the S_i (Whittemore 1996). For example, one can have $S_i = aS_i' + (1 - a)\, S_i''$, where S_i' and S_i'' are two different scoring functions and a is a free parameter to be fitted. For example, with sib pairs, introducing an extra df can allow us to have a full model for the sharing probabilities, instead of the additive model. It is emphasized that fitting these more complicated models does not require multipoint calculations that are different from or additional to those currently performed by GENEHUNTER. Hence the extra computational cost will be modest. Currently, we feel that the addition of more parameters is more important at a later stage, after linkage has been detected. Then data on additional markers in the region of interest would probably have been collected, and there might even be data on additional pedigrees. Having a model with one or two extra parameters that fits the data better can increase the resolution for localization. However, at the stage of testing for linkage, there is much to be said for having a 1-df alternative. For example, with sib pairs, if the true deviations from the null are not very large, then, even if they do not satisfy the additive model exactly, a likelihood-ratio test based on the additive model, apart from being simpler, will probably have more power, in many circumstances, than one based on the full model with df somewhere between 1 and 2 (Holmans 1993). In general, with a limited amount of data, one can lose power even with a more correct model, because one has to pay the price for the additional df.

Discussion

The power of a linkage test based on IBD sharing depends on two factors. The first factor is the choice of the scoring function(s) and weighting factors. The second factor, which is the main focus of this paper, is how to appropriately evaluate the statistical significance, given that choice. Whittemore (1996) demonstrates the correspondence between the NPL score and the score statistic of model (1). This relationship can be used to argue that the test based on the NPL score is not really *model free*. We take advantage of the relationship and the special incomplete-data property of the model, to derive a test, based on the likelihood ratio, that is accurate in the sense that it does not have the tendency to severely overestimate the P value. Moreover, the model allows us to obtain a LOD-score curve that can be used for the localization of the gene after the detection of linkage.

The optimal scoring function(s) and weighting factors will, in general, depend on the mode of inheritance and the ascertainment scheme. Since, for a complex trait, the mode of inheritance is usually unknown, the general strategy should be to choose the scoring function and weighting factors so that they will be close to optimal for a wide range of plausible modes of inheritance. The choice of the scoring function is currently an active area of research (Whittemore and Halpern 1994). For pedigrees that have two or three affecteds, Teng and Siegmund (1997) have provided some useful results for

choosing both S_i and γ_i. However, how to choose the scoring functions and weighting factors when a data set consists of pedigrees of various sizes and structures remains a challenging problem.

Although we appreciate the importance of the scoring function and the weighting factors, on the basis of our limited experience, those choices have a smaller effect on the LOD scores than does the specification of the penetrances in a traditional parametric analysis. For example, with the Hanis et al. data, we have tried both scoring functions, $\mathbf{S}_{\text{pairs}}$ and $\mathbf{S}_{\text{all}}$, and have also tried varying the weights according to sibship sizes; and the results do not change by a very substantial amount. This may be mainly because the Hanis et al. data consist only of sibships. However, this is also at least partly due to the fact that the value of δ, a measure of the gene effect, is estimated, at every location, to maximize the likelihood/LOD score. One may think of the scoring function and the weighting factors as the model and think of δ as the parameter. The implication is that, with complex traits, many different allele-sharing models, given an appropriate choice of the parameter value, can fit the data reasonably well.

Appendix A

Consider a single pedigree i. Let $\mathrm{L}(w) = P(\text{data}\,|w, \mathrm{H}_0) = P(\text{data}\,|w)$, making the important assumption that the distribution of the data on all markers, *given* the inheritance vector at the gene location, no longer depends on the actual distribution of the inheritance vector. (For example, with a sib pair, given that we know which one of the four IBD sharing possibilities has occurred at the assumed gene location, the distribution of the data no longer depends on the actual sharing probabilities.) The application of Bayes's rule results in

$$\begin{aligned} P(w|\text{data}, \mathrm{H}_0) &= \frac{P(\text{data}|w)P(w|\mathrm{H}_0)}{\sum_\eta P(\text{data}|\eta)P(\eta|\mathrm{H}_0)} \\ &= \frac{\mathrm{L}(w)c_i}{\sum_\eta \mathrm{L}(\eta)c_i} \\ &= \frac{\mathrm{L}(w)}{W} \ , \end{aligned}$$

where $W = \sum_\eta \mathrm{L}(\eta)$. It follows that

$$\begin{aligned} \bar{S}_i &= \mathrm{E}(S_i|\text{data}, \mathrm{H}_0) \\ &= \sum_w S_i(w)P(w|\text{data}, \mathrm{H}_0) \\ &= \frac{\sum_w S_i(w)\mathrm{L}(w)}{W} \ . \end{aligned}$$

Under the alternative model,

$$\begin{aligned} P(\text{data}|\delta) &= \sum_w P(\text{data}|w)P(w|\delta) \\ &= \sum_w L(w)c_i\{1 + \delta\gamma_i[S_i(w) - \mu_i]/\sigma_i\} \\ &= c_i\left[\sum_w \mathrm{L}(w)\right] + c_i(\delta\gamma_i/\sigma_i)\left[\sum_w S_i(w)\mathrm{L}(w)\right] \\ &\quad - c_i(\delta\gamma_i/\sigma_i)\mu_i\left[\sum_w \mathrm{L}(w)\right] \\ &= c_i W[1 + (\delta\gamma_i/\sigma_i)(\bar{S}_i - \mu_i)] \\ &= \text{constant} \times [1 + (\delta\gamma_i/\sigma_i)(\bar{S}_i - \mu_i)] \ . \end{aligned}$$

Hence, under this alternative model—and for this alternative model only—the likelihood function with respect to the imperfect data depends only on $\bar{S}_i$. The overall log likelihood, $\ln[P(\text{all data}\,|\,\delta)]$, which takes all m pedigrees into account, is

$$\begin{aligned} \mathrm{l}(\delta) &= C + \sum_{i=1}^{m} \ln[1 + \delta\gamma_i(\bar{S}_i - \mu_i)/\sigma_i] \\ &= C + \sum_{i=1}^{m} \ln(1 + \delta\gamma_i\bar{Z}_i) \\ &= C + \sum_{i=1}^{m} \ln(1 + \delta T_i) \ , \end{aligned}$$

where C is the constant that does not depend on δ.

Appendix B

A modified version of GENEHUNTER is available via anonymous ftp at galton.uchicago.edu in the /pub/kong directory. A tarfile containing a compiled version for a SUN SPARC running SUNOS 2.x is supplied, as well as the complete (modified) sources.

One additional command is implemented in the modified version, the "kac" command. This command, which can be issued only after a successful "scan," pro-

duces an output file with a listing of the various statistics (NPL score, Z_{lr}, and LOD[1]) described in this paper, for each GENEHUNTER evaluation location. Currently, just like GENEHUNTER, only $\mathbf{S}_{pairs}$ or $\mathbf{S}_{all}$ with equal weighting is allowed. The on-line help file and user manual have been modified to include instructions on the use of this new "kac" command.

Acknowledgments

This research was facilitated in part by National Institutes of Health grants R01-GM46800, DK-20595, DK-47481, DK-47486, DK-47487, and DK-47494. The authors thank Mike Frigge for his help in modifying the GENEHUNTER program and thank David Wallace and Peter McCullagh for their valuable comments and suggestions.

References

Cox DR, Hinkley DV (1994) Theoretical statistics. Chapman & Hall, London

Hanis CL, Boerwinkle E, Chakraborty R, Ellsworth DL, Concannon P, Stirling B, Morrison VA, et al (1996) A genome-wide search for human non-insulin-dependent (type 2) diabetes genes reveals a major susceptibility locus on chromosome 2. Nat Genet 13:161–166

Holmans P (1993) Asymptotic properties of affected-sib-pair linkage analysis. Am J Hum Genet 52:362–374

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347–1363

Lander ES, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241–247

Risch N (1990*a*) Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46:222–228

——— (1990*b*) Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 46:229–241

SAGE (1994) Statistical analysis for genetic epidemiology, release 2.2. Computer package available from the Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland

Teng J, Siegmund D (1997) Combining information within and between pedigrees for mapping complex traits. Am J Hum Genet 60: 979–992

Weeks DE, Lange K (1988) The affected-pedigree-member method of linkage analysis. Am J Hum Genet 42:315–326

Whittemore AS (1996) Genome scanning for linkage: an overview. Am J Hum Genet 59:704–716

Whittemore AS, Halpern J (1994) A class of tests of linkage using affected pedigree members. Biometrics 50:118–127